

Concentration inequalities of the cross-validation estimate for stable predictors

Matthieu Cornec

November 24, 2010

Abstract

In this article, we derive concentration inequalities for the cross-validation estimate of the generalization error for stable predictors in the context of risk assessment. The notion of stability has been first introduced by [DEWA79] and extended by [KEA95], [BE01] and [KUNY02] to characterize class of predictors with infinite VC dimension. In particular, this covers k -nearest neighbors rules, bayesian algorithm ([KEA95]), boosting, ... General loss functions and class of predictors are considered. We use the formalism introduced by [DUD03] to cover a large variety of cross-validation procedures including leave-one-out cross-validation, k -fold cross-validation, hold-out cross-validation (or split sample), and the leave- v -out cross-validation.

In particular, we give a simple rule on how to choose the cross-validation, depending on the stability of the class of predictors. In the special case of uniform stability, an interesting consequence is that the number of elements in the test set is not required to grow to infinity for the consistency of the cross-validation procedure. In this special case, the particular interest of leave-one-out cross-validation is emphasized.

Keywords: Cross-validation, stability, generalization error, concentration inequality, optimal splitting, resampling.

1 Introduction and motivation

One of the main issue of pattern recognition is to create a predictor (a regressor or a classifier) which takes observable inputs in order to predict the unknown nature of an output. Formally, a predictor φ is a measurable map from some measurable space \mathcal{X} to some measurable space \mathcal{Y} . When \mathcal{Y} is a countable set (respectively \mathbb{R}^m), the predictor is called a classifier (respectively a regressor). The strategy of *Machine Learning* consists in building a learning algorithm Φ from both a set of examples and a class of methods. Typical class of methods are empirical risk minimization or k -nearest neighbors rules. The set of examples consists in the measurement of n observations $(x_i, y_i)_{1 \leq i \leq n}$. Thus, formally, Φ is a measurable map from $\mathcal{X} \times \cup_n (\mathcal{X} \times \mathcal{Y})^n$ to \mathcal{Y} . One of the main issue of *Statistical Learning* is to analyze the performance of a learning algorithm in a probabilistic setting. $(x_i, y_i)_{1 \leq i \leq n}$ are supposed to be observations from n independent and identically distributed (i.i.d.) random variables $(X_i, Y_i)_{1 \leq i \leq n}$ with unknown distribution \mathbb{P} . $(X_i, Y_i)_{1 \leq i \leq n}$ is denoted \mathcal{D}_n in the following and called the learning set. In order to analyze the performance, it is usual to consider the conditional risk of a machine learning Φ denoted \tilde{R}_n , so called the generalization error. It is defined by the conditional expectation of $L(Y, \Phi(X, \mathcal{D}_n))$ given \mathcal{D}_n where $(X, Y) \sim \mathbb{P}$ is a random variable independent of \mathcal{D}_n , i.e. $\tilde{R}_n := \mathbb{E}_{X, Y}(L(Y, \Phi(X, \mathcal{D}_n)) | \mathcal{D}_n)$ with L a cost function from $\mathcal{Y}^2 \rightarrow \mathbb{R}_+$. Notice that \tilde{R}_n is a random variable measurable with respect to \mathcal{D}_n .

An important question is: the distribution \mathbb{P} of the generating process being unknown, can we estimate how good a predictor trained on a learning set of size n is? In other words, can we approximate the generalization error \tilde{R}_n ? This fundamental statistical problem is referred to "choice and assessment of statistical predictions" [STO74]. Many estimates have been proposed. Quoting [HTF01]: *Probably the simplest and most widely used method for estimating prediction error is cross-validation.*

The cross-validation procedures include leave-one-out cross-validation, k -fold cross-validation, hold-out cross validation (or split sample), leave- v -out cross-validation (or Monte Carlo cross-validation or bootstrap cross-validation). With the exception of [BUR89], theoretical investigations of multifold cross-validation procedures have first concentrated on linear models ([Li87]; [SHAO93]; [ZHA93]). Results of [DGL96] and [GYO02] are discussed in Section 3. The first finite sample results are due to Wagner and Devroye [DEWA79] and concern k -local rules algorithms under leave-one-out and hold-out cross-validation. More recently, [HOL96, HOL96bis] derived finite sample results for v -out cross-validation, k -fold cross-validation, and leave-one-out cross-validation for Empirical Risk Minimization (ERM) over a class of predictors with finite Vapnik-Chervonenkis-dimension (VC-dimension) in the realisable case (the generalization error is equal to zero). [BKL99] have emphasized when k -fold can beat v -out cross-validation in the particular case of k -fold predictor. [KR99] has extended such results in the case of stable algorithms for the leave-one-out cross-validation procedure. [KEA95] also derived results for hold-out cross-validation for ERM, but their arguments rely on the traditional notion of VC-dimension. In the particular case of ERM over a class of predictors with finite VC-dimension but with general cross-validation procedures, we derived derived probability upper bounds in chapter 1: we denote by p_n the percentage of elements in the test sample. In the sequel, we will denote by \hat{R}_{CV} the cross-validation estimator. For empirical risk minimizers over a class of predictors with finite VC-dimension V_C , to be defined below, we obtained the following concentration inequality. For all $\varepsilon > 0$, we have

$$\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon),$$

with

- $B(n, p_n, \varepsilon) = 5(2n(1 - p_n) + 1)^{\frac{4V_C}{1-p_n}} \exp(-\frac{n\varepsilon^2}{64}),$
- $V(n, p_n, \varepsilon) = \min \left(\exp(-\frac{2np_n\varepsilon^2}{25}), \frac{16}{\varepsilon} \sqrt{\frac{V_C(\ln(2(1 - p_n) + 1) + 4)}{n(1 - p_n)}} \right).$

Unfortunately, many popular predictors, including k -nearest neighbors rules, do not satisfy this property. Moreover, these bounds obtained are called "sanity check bounds" since they are not better than classical Vapnik-Chernovenkis's bounds.

To avoid the traditional analysis in the VC framework, notions of stability have been intensively worked through in the late 90's [KEA95], [BE01], [BE02], [KUT02], and [KUNIY02]. The object of stability framework is the learning algorithm rather than the space of classifiers. The learning algorithm is a map (effective procedure) from data sets to classifiers. An algorithm is stable at a learning set \mathcal{D}_n if changing one point in \mathcal{D}_n yields only a small change in the output hypothesis. The attraction of such an approach is that it avoids the traditional notion of VC-dimension, and allows to focus on a wider class of learning algorithms than empirical risk minimization. For example, this approach provides generalization error bounds for regularization-based learning algorithms that have been difficult to analyze within the VC framework such as boosting. As a motivation, we quote the following list of algorithms satisfying stability properties: regularization networks, ERM, k -nearest rules, boosting.

Algorithmic stability was first introduced by [DEWA79]. [?] argued that unstable weak learners benefit from randomization algorithms such as bagging. [KR99] considered both algorithmic stability and the weaker related notion of error stability. They proved bounds on the error of cross-validation estimates of generalization error, but their arguments rely on VC theory. [BE01, BE02] proved that an algorithm which is stable everywhere has low generalization error; their proof does not make any reference to VC-dimension. They showed that regularization networks are stable. In [KUNIY02], at least ten different notions were examined. In particular, they introduced a probabilistic notion of change-one stability called Cross-Validation stability or CV stability. This was shown to be necessary and sufficient for consistency of ERM in the Probably Approximately Correct (PAC) Model of [VAL84].

The goal of this paper is to obtain exponential bounds to fill the chart 1 where possible bounds are missing (up to our knowledge).

	leave-one-out	hold-out	k-fold	ν -out
ERM with				
finite VC-dimension	Kearns, Holden, Cornec	Holden, Cornec	Holden, Cornec	Cornec
hypothesis stability	Devroye and W	Devroye and W	×	×
error stability				
with finite VC dimension	Kearns	Kearns	×	×
uniform stability	Bousquet and E.	×	×	×
strong hypothesis	Kutin and N	×	×	×
weak stability	×	×	×	×

Table 1: Missing bounds × to find

The goal of this article is also to show that cross-validation is still consistent for stable predictors. As a consequence, we will emphasize the role played by cross-validation: it can be a consistent estimate of the generalisation error when the training error defined by $\hat{R}_n := \frac{1}{n} \sum_{i=1}^n L(Y_i, \phi(X_i, \mathcal{D}_n))$ is not. Indeed, for stable predictors, the training error can be arbitrarily poor: for example, the training error for 1-nearest neighbor is equal to zero whatever the generalisation error may be.

We introduce our **main result**¹. Suppose that the cross-validation is symmetric -i.e. the probability of a observation to be in the training set is independent of its index- and that the number of elements in the test set is constant and equal to np_n with p_n the percentage of elements in the test set. All the bounds of the following form $\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon + \dots) \leq B(n, p_n, \varepsilon) + V(n, p_n, \varepsilon)$.

Under certain stability conditions -satisfied for example by Empirical Risk Minimisers (ERM) or Adaboost-, we have for all $\varepsilon \geq 0$,

$$\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon + 2\lambda p_n) \leq 2 \exp(-2np_n\varepsilon^2) + \delta_{n,p_n}$$

¹accurate inequalities can be found in section 3

with δ_{n,p_n} and λ a non-negative real numbers. For classical algorithms, we have in mind that $\delta_{n,p_n} = O_n(p_n \exp(-n(1 - p_n)))$. λ is in fact a Lipschitz coefficient with respect to the total variation and can be interpreted as a stability factor: the smaller λ is, the more stable the learning algorithm is. Furthermore, if the learning algorithm satisfies a stronger stability condition (for example Adaboost or regularization networks), we obtain

$$\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + 2\lambda p_n) \leq 4(\exp(-\frac{\varepsilon^2}{8(18\lambda)^2 n p_n^2}) + \frac{n}{9\lambda p_n} \delta'_{n,p_n})$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n,1/n}$. For the latter, it is thus not required that the number of elements in the test set grows to infinity for the consistency of the cross-validation to hold.

Using these probability bounds, we can then deduce that the expectation between the generalization error and the cross-validation error $\mathbb{E}_{\mathcal{D}_n}|\hat{R}_{CV} - \tilde{R}_n|$ is of order $O_n((\lambda/n)^{1/3})$. As far as the expectation $\mathbb{E}_{\mathcal{D}_n}|\hat{R}_{CV} - \tilde{R}_n|$ is concerned, we can define a splitting rule in the general setting: the percentage of elements p_n^* in the test set should be proportional to $(1/\lambda^2 n)^{1/3}$, i.e. the less stable (i.e. λ large) the learning algorithm is, the smaller the test set in the cross-validation should be. Furthermore, if the learning algorithm satisfies a stronger stability condition (for example Adaboost or regularization networks), we also have $\mathbb{E}_{\mathcal{D}_n}|\hat{R}_{CV} - \tilde{R}_n| = O_n(\lambda/\sqrt{n})$ and the leave-one-out cross-validation (i.e. $p_n^* = 1/n$) is preferred for n large enough.

The paper is organized as follows. In the next section, we recall the main notations and definitions of cross-validation as introduced in chapter 1. We also introduce notations to unify the main notions of stability. Finally, in Section 3, we introduce our results in terms of probability upperbounds. We also prove that many traditionnal methods satisfy our generalized notion of stability (lasso,...,adaboost, k-nearest neighbors).

2 Notations and definitions

In the following, we follow the notations of cross-validation introduced in chapter 1.

2.1 Cross-validation

We will consider the following shorter notations inspired by the literature on empirical processes. In the sequel, we will denote $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and $(Z_i)_{1 \leq i \leq n} := ((X_i, Y_i))_{1 \leq i \leq n}$ the learning set. For a given loss function L and a given class of predictors \mathcal{G} , we define a new class \mathcal{F} of functions from \mathcal{Z} to \mathbb{R}_+ by $\mathcal{F} := \{\psi \in \mathbb{R}_+^{\mathcal{Z}} | \psi(Z) = L(Y, \phi(X)), \phi \in \mathcal{G}\}$. For a machine learning Φ , we have the natural definition $\Psi(Z, \mathcal{D}_n) := L(Y, \Phi(X, \mathcal{D}_n))$. With these notations, the conditional risk \tilde{R}_n is the expectation of $\Psi(Z, \mathcal{D}_n)$ with respect to \mathbb{P} conditionally on \mathcal{D}_n : $\tilde{R}_n := \mathbb{E}_Z[\Psi(Z, \mathcal{D}_n) | \mathcal{D}_n]$ with $Z \sim \mathbb{P}$ independent of \mathcal{D}_n . In the following, if there is no ambiguity, we will also allow the following notation $\psi(X, \mathcal{D}_n)$ instead of $\Psi(X, \mathcal{D}_n)$.

To define the accurate type of cross-validation procedure, we introduce binary vectors. Let $V_n = (V_{n,i})_{1 \leq i \leq n}$ be a vector of size n . V_n is a binary vector if for all $1 \leq i \leq n$, $V_{n,i} \in \{0, 1\}$ and if $\sum_{i=1}^n V_{n,i} \neq 0$. Consequently, we can define the subsample associated with it, $\mathcal{D}_{V_n} := \{Z_i \in \mathcal{D}_n | V_{n,i} = 1, 1 \leq i \leq n\}$. We define a weighted empirical measure on \mathcal{Z}

$$\mathbb{P}_{n,V_n} := \frac{1}{\sum_{i=1}^n V_{n,i}} \sum_{i=1}^n V_{n,i} \delta_{Z_i},$$

with δ_{Z_i} the Dirac measure at $\{Z_i\}$. We also define a weighted empirical error $\mathbb{P}_{n,V_n} \psi$ where $\mathbb{P}_{n,V_n} \psi$ stands for the usual notation of the expectation of ψ with respect to \mathbb{P}_{n,V_n} . For $\mathbb{P}_{n,1_n}$, with 1_n the

binary vector of size n with 1 at every coordinate, we will use the traditional notation \mathbb{P}_n . For a predictor trained on a subsample, we define

$$\psi_{V_n}(\cdot) := \Psi(\cdot, \mathcal{D}_{V_n}).$$

With the previous notations, notice that the predictor trained on the learning set $\psi(\cdot, \mathcal{D}_n)$ can be denoted by $\psi_{1_n}(\cdot)$. We will allow the simpler notation $\psi_n(\cdot)$. The learning set is divided into two disjoint sets: the training set of size $n(1 - p_n)$ and the test set of size np_n , where p_n is the percentage of elements in the test set. To represent the training set, we define V_n^{tr} a random binary vector of size n independent of \mathcal{D}_n . V_n^{tr} is called the training vector. We define the test vector by $V_n^{ts} := 1_n - V_n^{tr}$ to represent the test set.

The distribution of V_n^{tr} characterizes all the cross-validation procedures described in the previous section (see e.g. chapter 1). Using our notations, we can now define the cross-validation estimator.

Definition 1 (Cross-validation estimator) *With the previous notations, the generalized cross-validation error of ψ_n denoted $\widehat{R}_{CV}(\psi_n)$ is defined by*

$$\widehat{R}_{CV}(\psi_n) := \mathbb{E}_{V_n^{tr}, V_n^{ts}} (\psi_{V_n^{tr}}).$$

We will give here an example of distributions of V_n^{tr} to illustrate we retrieve cross-validation procedures described previously. Leave- v -out cross-validation is an elaborate and expensive version of cross-validation. This procedure divides the data into two sets: the training set of size $n - v$ and the test set of size v . It then produces a predictor by training on the training set and testing on the remaining test set. This is repeated for all possible subsamples of v cases, and the observed errors are averaged to form the leave- v -out estimate. Denote by $(\xi_{n,i}^v)_{1 \leq i \leq \binom{n}{v}}$ the family of binary vectors of size n such that $\sum_{i=1}^{\binom{n}{v}} \xi_{n,i}^v = n - v$.

Example 2 (Leave- v -out cross-validation)

$$\begin{aligned} \Pr(V_n^{tr} = \xi_{n,1}^v) &= \frac{1}{\binom{n}{v}} \\ \Pr(V_n^{tr} = \xi_{n,2}^v) &= \frac{1}{\binom{n}{v}} \\ &\dots \\ \Pr(V_n^{tr} = \xi_{n,\binom{n}{v}}^v) &= \frac{1}{\binom{n}{v}}. \end{aligned}$$

For other examples, see chapter one.

2.2 Definitions and notations of stability

The basic idea is that an algorithm is stable at a training set \mathcal{D}_n if changing one point in \mathcal{D}_n yields only a small change in the output hypothesis. Formally, a learning algorithm maps a weighted training set into a predictor space. Thus, stability can be translated into a Lipschitz condition for this mapping with high probability.

To be more formal, following [KUNIY02], we define a distance between two weighted empirical errors.

Let \mathbb{P}_{n,V_n} and \mathbb{P}_{n,U_n} be two empirical measures on \mathcal{Z} with respect to the binary vectors V_n and U_n . We do not assume their support to be equal. The distance between them is defined as their total variation, i.e. the number of points they do not have in common

$$\|\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n}\| = \sup_{A \in \mathcal{P}(\mathcal{Z})} |(\mathbb{P}_{n,U_n} - \mathbb{P}_{n,V_n})(A)|.$$

Example 3 In the case of leave-one-out (i.e. $\sum_{i=1}^n U_{n,i} = n - 1$), we have

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2}{n}.$$

In the case of leave- ν -out, we get

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = \frac{2\nu}{n}.$$

In the general setting, it follows that

$$||\mathbb{P}_{n,U_n} - \mathbb{P}_n|| = 2p_n.$$

At least, we need a distance d on the set \mathcal{F} . Let us quote three important examples. Let $\psi_1, \psi_2 \in \mathcal{F}$. The uniform distance is defined by: $d_\infty(\psi_1, \psi_2) = \sup_{Z \in \mathcal{Z}} |\psi_1(Z) - \psi_2(Z)|$, the L_1 -distance by: $d_1(\psi_1, \psi_2) = \mathbb{P}|\psi_1 - \psi_2|$, the error-distance $d_e(\psi_1, \psi_2) = |\mathbb{P}(\psi_1 - \psi_2)|$. It is important to notice that what matters here is not an absolute distance between the original class of predictors \mathcal{G} seen as functions but the distance with the respect to the loss or/and the distribution \mathbb{P} . In particular, for the L_1 -distance, we do not care about the behavior of the original predictors φ_1 and φ_2 outside the support of \mathbb{P} . At last, notice that we always have $d_e \leq d_1 \leq d_\infty$.

We are now in position to define the different notions of stability of a learning algorithm which cover notions introduced by [KUNIY02]. We begin with the notion of weak stability. In essence, it says that for any given resampling vectors, the distance between two predictors is controlled with high probability by the distance between the resampling vectors.

Definition 4 (Weak stability) Let $\alpha, \lambda, (\delta_{n,p_n})_{n,p_n}$ be nonnegative real numbers. A learning algorithm Ψ is said to be weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable if for any training vector U_n whose sum is equal to $n(1 - p_n)$:

$$\Pr(d(\psi_{U_n}, \psi_n) \geq \lambda ||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha) \leq \delta_{n,p_n}.$$

Notice that in the former definition \Pr stands for $\mathbb{P}^{\otimes n}$. Indeed, ψ_n is trained with n observations, drawn independently from \mathbb{P} . A stronger notion is to consider ψ_n trained with $n - 1$ observations drawn independently from \mathbb{P} and an additionnal general observation z . We consider the stronger notion of strong stability. As a motivation, notice that algorithms such as Empirical Risk Minimization with finite VC dimension ([KUNIY02]) satisfies this property.

Definition 5 (Strong stability) Let $z \in \mathcal{Z}$. Let $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ be a learning set. Let $\lambda, (\delta_{n,p_n})_{n,p_n}$ be nonnegative real numbers. A learning algorithm Ψ is said to be strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable if for any training vector U_n whose sum is equal to $n(1 - p_n)$

$$\Pr(d(\psi_{U_n}, \psi_n) \geq \lambda ||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha) \leq \delta_{n,p_n}.$$

What we have in mind for classical algorithms is $\delta_{n,p_n} = O_n(p_n \exp(-n(1 - p_n)))$. We can state the last definition in other words. Let V_n^{tr} be a training vector with distribution \mathbb{Q} such that the number of elements in the training set is constant and equal to $n(1 - p_n)$. Notice then that the former definition also implies that $\sup_{U_n \in \text{support}(\mathbb{Q})} \mathbb{P}(\frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha} \geq \lambda) \leq \delta_{n,p_n}$, where $\text{support}(\mathbb{Q})$ stands for the support of \mathbb{Q} . The previous notion stands for any U_n having the same support of \mathbb{Q} . A stronger hypothesis would be that the previous probability stands uniformly over U_n in $\text{support}(\mathbb{Q})$. This leads formally to the notion of cross-validation stability. As a motivation, notice that algorithms such as Lasso ([BTW07]) satisfies this property. To be more accurate, we define

Definition 6 (Cross-validation weak stability) Let $\mathcal{D}_n = (Z_i)_{1 \leq i \leq n}$ a learning set. Let V_n^{tr} a training vector with distribution \mathbb{Q} . Let $\lambda, (\delta_{n,p_n})_{n,p_n}$ be nonnegative real numbers. A learning algorithm Ψ is said to be weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable if it is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable and if:

$$\Pr(\sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{||\mathbb{P}_{n,U_n} - \mathbb{P}_n||^\alpha} \geq \lambda) \leq \delta_{n,p_n}.$$

As before, we also define the following stronger notion

Definition 7 (Cross-validation strong stability) Let $z \in \mathcal{Z}$. Let $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{z\}$ a learning set. Let V_n^{tr} be a cross-validation vector with distribution \mathbb{Q} . A learning algorithm Ψ is said to be strongly $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable if it is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable and if:

$$\Pr\left(\sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{\|\mathbb{P}_{n,U_n} - \mathbb{P}_n\|^\alpha} \geq \lambda\right) \leq \delta_{n,p_n}.$$

Remark 8 If the cardinal of the support of \mathbb{Q} is denoted $\kappa(n)$, then a learning algorithm which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ -stable is also strong $(\lambda, (\kappa(n)\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ -stable.

At last, we consider the special important case when $\delta_{n,p_n} = 0$. This is the case in particular for regularization networks ([BE01]).

Definition 9 (Sure stability) Notice that when $\delta_{n,p_n} = 0$, the two notions coincides and are called **sure stability**.

As an example of strong stability, we develop the description of [FRE95] who introduced the algorithm.

Example 10 (Adaboost) We give an initial distribution p^1 and let $w^1 = p^1$ and $Z_1 = 1$. Let Φ be a learning algorithm. Let T the number of rounds.

For each $t = 1 \dots T$:

1. Train the learning algorithm Φ on the learning set with distribution $p^{(t)}$. The predictor obtained is denoted by $\varphi^{(t)}$.
2. For each i , let $a_i^t = |\varphi^{(t)}(x_i) - y_i|$, the error of $\varphi^{(t)}$ on instance i .
3. Let $\varepsilon_t = \sum_{i=1}^m p^t a_i^t$, the error rate of $\varphi^{(t)}$ with respect to $p^{(t)}$.
4. Let $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$ and let $\alpha_t = \ln(1/\beta_t)$
5. reweight the data: for all i , let $w_i^{(t+1)} = w_i^{(t)} \beta_t^{1-a_i^t}$.
6. Normalize the distribution: let $Z_{t+1} = \sum_{i=1}^m w_i^{(t+1)}$ and $p_i^{(t+1)} = w_i^{(t+1)} / Z_{t+1}$

The final output is $H_T(x) = \sum_{s=1}^T \alpha_s \varphi^{(s)}(x)$.

[KUNY01] shows that under certain hypotheses, Adaboost is strongly stable: suppose the learner Φ - $(\lambda, 0, d_\infty)$ stable and other regularity assumptions, then Adaboost with T rounds is strong $(\lambda^*, (\delta_{n,p_n}^*)_{n,p_n}, d_\infty)$ stable for some λ^* and δ_{n,p_n}^* .

We give now an example that is surely stable introduced in [BE01].

Example 11 (Regularization networks) Regularization networks are attractive for their links with Support Vector Machines and their Bayesian interpretation. This learning algorithm consists in finding a function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ in a space H which minimizes the following functional:

$$A(\varphi) = \frac{1}{n} \sum_{i=1}^n (Y_i - \varphi(X_i))^2 + \lambda \|\varphi\|_H^2,$$

with $\|\varphi\|_H$ the L_2 norm in the space H . H is chosen to be a reproducing kernel Hilbert Space (rkhs) with kernel k . k is supposed to be a symmetric function $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. In particular, we have the following property (for a detailed introduction of rkhs, see [ATE92])

$$|f(x)| \leq \|f\|_H \|k\|_H.$$

We slightly adapt the proof in [BE01] to show that a regularization network is surely stable:

Theorem 12 If Ψ is a regularization network such that $\|k\|_H \leq \kappa$ and $(y - \varphi(x))^2 \leq M$, then Ψ is $\frac{4M\kappa^2}{n\lambda}$ -surely stable with respect to the distance d_∞ .

Proof

Define $A^i(\varphi) := \frac{1}{n-1} \sum_{j \neq i} (Y_j - \varphi(X_j))^2 + \lambda \|\varphi\|_H^2$ and $\mathcal{D}_n^i := \mathcal{D}_n \setminus \{(X_i, Y_i)\}$. $\varphi_{\mathcal{D}_n^i}$ is the minimizer of A^i over H whereas $\varphi_{\mathcal{D}_n}$ is the minimizer of A . Denote $g := \varphi_{\mathcal{D}_n^i} - \varphi_{\mathcal{D}_n}$.

For $t \in [0, 1]$, we have $A(\varphi_{\mathcal{D}_n}) - A(\varphi_{\mathcal{D}_n} + tg)$ is equal to

$$\frac{-2t}{n(n-1)} \sum_{j \neq i} (n-1)(\varphi_{\mathcal{D}_n}(x_j) - y_j)g(x_j) - \frac{2t}{n(n-1)} (n-1)(\varphi_{\mathcal{D}_n}(x_i) - y_i)g(x_i) - 2t\lambda \langle \varphi_{\mathcal{D}_n}, g \rangle_H + t^2 B(g)$$

with $B(g)$ the factor of t^2 .

In the same way, we get that $A^i(\varphi_{\mathcal{D}_n^i}) - A^i(\varphi_{\mathcal{D}_n^i} - tg)$ is equal to

$$\frac{2t}{n(n-1)} \sum_{j \neq i} (n-1)(\varphi_{\mathcal{D}_n^i}(x_j) - y_j)g(x_j) + \frac{2t}{n(n-1)} \sum_{j \neq i} (\varphi_{\mathcal{D}_n^i}(x_j) - y_j)g(x_j) + 2t\lambda \langle \varphi_{\mathcal{D}_n^i}, g \rangle_H + t^2 B^n(g).$$

By definition of A and A^i , we have $A(\varphi_{\mathcal{D}_n}) - A(\varphi_{\mathcal{D}_n} + tg) \leq 0$ and $A^i(\varphi_{\mathcal{D}_n^i}) - A^i(\varphi_{\mathcal{D}_n^i} - tg) \leq 0$. Thus, we get by summing these two inequalities, dividing by $\frac{2t}{n(n-1)}$ and making $t \rightarrow 0$.

$$\sum_{j \neq i} (n-1)g^2(x_j) + \sum_{j \neq i} \left[(\varphi_{\mathcal{D}_n^i}(x_j) - y_j)g(x_j) - (\varphi_{\mathcal{D}_n}(x_i) - y_i)g(x_i) \right] + n(n-1)\|g\|_H^2 \leq 0$$

which leads to

$$n(n-1)\|g\|_H^2 \leq \sum_{j \neq i} \left[(\varphi_{\mathcal{D}_n}(x_i) - y_i)g(x_i) - (\varphi_{\mathcal{D}_n^i}(x_j) - y_j)g(x_j) \right] \leq 2(n-1)\sqrt{M}\kappa\|g\|_H$$

by assumptions.

Thus, we have

$$\|g\|_H \leq 2(n-1)\sqrt{M}\kappa\|g\|_H/n\lambda$$

and also, for all x, y

$$|(\varphi_{\mathcal{D}_n}(x) - y)^2 - (\varphi_{\mathcal{D}_n^i}(x) - y)^2| \leq 2\sqrt{M}|\varphi_{\mathcal{D}_n}(x) - \varphi_{\mathcal{D}_n^i}(x)| \leq 4M\kappa^2/n\lambda.$$

□

Another popular example is given by the k -nearest neighbors which are strongly stably with respect to d_1 .

Example 13 (k-nearest neighbors) In the k -nearest rule, the machine learning is a function of X and of the k nearest observations to X from (X_1, \dots, X_n) and of the corresponding (Y_1, \dots, Y_n) . Because there may be ties in determining the k nearest neighbors, we use an independent sequence (Z, Z_1, \dots, Z_n) of i.i.d uniform random variables in $[0, 1]$. X_j is nearer X_i to X if:

1. $\|X_j - X\| < \|X_i - X\|$ or
2. $\|X_j - X\| = \|X_i - X\|$ and $|Z_j - Z| < |Z_i - Z|$, or
3. $\|X_j - X\| = \|X_i - X\|$ and $Z_j = Z_i$ and $j < i$.

The last event does not count since its has zero probability.

Denote γ_d the maximum number of distinct points in \mathbb{R}^d that share the same nearest neighbor. It can be shown that $\gamma_d \leq 3^d - 1$ and other lower and upper bounds can be found in [ROG63]. Recall the following lemma from [DEWA79]: suppose $(X_1, Z_1), \dots, (X_n, Z_n)$ is the sequence obtained from the data by omitting the Y_1, \dots, Y_n . If, for each j , the nearest neighbor to (X_j, Z_j) is found from $(X_1, Z_1), \dots, (X_{j-1}, Z_{j-1}), (X_{j+1}, Z_{j+1}), \dots, (X_n, Z_n)$. Then no point (X_i, Y_i) can be the nearest neighbors to more than $\gamma_d + 2$ of the remaining points.

We can derive the next result following the proofs in [DEWA79].

Theorem 14 Let $\mathcal{D}_n := ((X_1, Z_1, Y_1), \dots, (x, z, y))$ be a learning set. Suppose Φ is a k local rule. Then we have for all $\varepsilon > 0$,

$$\Pr(E_{X,Y,Z} | L(Y, \Phi((X, Z), \mathcal{D}_n)) - L(Y, \Phi((X, Z), \mathcal{D}_n^i))| \geq \varepsilon) \leq 6 \exp\left(\frac{-(n-1)\varepsilon^3}{54k(\gamma_d + 2)}\right)$$

with $\mathcal{D}_n^i := \mathcal{D}_n \setminus \{(X_i, Y_i, Z_i)\}$ and i a fixed index.

It says that the k nearest rule satisfies strong stability property with respect d_1 and $\|\mathbb{P}_{n,U_n} - \mathbb{P}_n\|^\alpha$ with $\alpha < 1/3$.

Proof

Consider one local rule first.

Let m be an integer. Consider an independent identically distributed ghost sample

$$((X_{n+1}, Y_{n+1}, Z_{n+1}), \dots, (X_{n+m}, Y_{n+m}, Z_{n+m})).$$

Denote $\mathcal{T}_{n+m} := ((X_1, Y_1, Z_1), \dots, (X_{n+m}, Y_{n+m}, Z_{n+m}))$ and $\mathcal{T}_{n+m}^j := \mathcal{T}_{n+m} \setminus \{(X_j, Y_j, Z_j)\}$.

- $L_1 := E_{X,Y,Z} | L(Y, \Phi((X, Z), \mathcal{D}_n)) - L(Y, \Phi((X, Z), \mathcal{D}_n^i)) |$
- $L_2 := \frac{1}{m} \sum_{j=1}^m |L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n)) - L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n^i))|$
- $L_3 := \frac{1}{m} \sum_{j=1}^m |L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n)) - L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{T}_{n+m}^{n+j}))|$
- $L_4 := \frac{1}{m} \sum_{j=1}^m |L(Y_j, \Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n^i)) - L(Y_{n+j}, \Phi((X_{n+j}, Z_{n+j}), \mathcal{T}_{n+m}^{n+j}))|$.

We have

$$\Pr(L_1 \geq 3\varepsilon) \leq \Pr(L_1 - L_2 \geq \varepsilon) + \Pr(L_3 \geq \varepsilon) + \Pr(L_4 \geq \varepsilon).$$

By Hoeffding's inequality we have $\Pr(L_1 - L_2 \geq \varepsilon) \leq \exp(-2m\varepsilon^2)$.

Now we get for the second term

$$\begin{aligned} \Pr(L_3 \geq \varepsilon) &\leq \Pr\left(\frac{1}{m} \sum_{j=1}^m 1_{\Phi((X_{n+j}, Z_{n+j}), \mathcal{D}_n) \neq \Phi((X_{n+j}, Z_{n+j}), \mathcal{T}_{n+m}^{n+j})} \geq \varepsilon\right) \\ &\leq \Pr\left(\frac{1}{m} \sum_{j=1}^m 1_{A(n+j)} \geq \varepsilon\right), \end{aligned}$$

with $A(n+j)$ the event that the nearest neighbor of (X_{n+j}, Z_{n+j}) from \mathcal{T}_{n+m}^{n+j} is attained in the ghost sample $\mathcal{T}_{n+m}^{n+j} \setminus \mathcal{D}_n$.

From [DEWA79], we have, if $(\gamma_d + 2)m < (n + m)\varepsilon/2$,

$$\Pr\left(\frac{1}{m} \sum_{j=1}^m 1_{A(n+j)} \geq \varepsilon\right) \leq 2 \exp(-2m(\varepsilon/2)^2)$$

In the same way, we find that $\Pr(L_4 \geq \varepsilon) \leq 2 \exp(-2m(\varepsilon/2)^2)$ if $(\gamma_d + 2)m < (n - 1 + m)\varepsilon/2$

Taking $m = \frac{(n-1)\varepsilon}{\gamma_d+2}$, we obtain

$$\Pr(L_3 \geq \varepsilon) \leq 2 \exp\left(\frac{-(n-1)\varepsilon^3}{2(\gamma_d+2)}\right)$$

and $\Pr(L_3 \geq \varepsilon) \leq 2 \exp\left(\frac{-(n-1)\varepsilon^3}{2(\gamma_d+2)}\right)$.

For an arbitray k , it is sufficient to replace $(\gamma_d + 2)$ by $k(\gamma_d + 2)$.

□

A last popular example is given by the Lasso which is strongly stable with respect to d_1 .

Example 15 (Lasso) We follow [BTW07] who defines Lasso-type methods in the following way. Let $((X_1, Y_1), \dots, (X_n, Y_n))$ be a sample of i.i.d. pairs distributed as $(X, Y) \in (\mathcal{X}, \mathbb{R})$, where \mathcal{X} is a borel subset of \mathbb{R}^d . We denote by μ the distribution of X on \mathcal{X} . Let $f(X) = \mathbb{E}(Y|X)$ be the unknown regression function and $\mathcal{F}_M = \{f_1, \dots, f_M\}$ be a dictionary of real-valued functions f_j that are defined on \mathcal{X} . We use a data dependent l_1 -penalty. Formally, for any $\lambda = (\lambda_1, \dots, \lambda_M) \in \mathbb{R}^M$, define $f_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$. Then the penalized least squares estimator of λ is

$$\hat{\lambda} = \arg \min \left\{ 1/n \sum_{i=1}^n (Y_i - f_\lambda(X_i))^2 + \text{pen}(\lambda) \right\}$$

where

$$\text{pen}(\lambda) = 2 \sum_{j=1}^M \omega_{n,j} |\lambda_j| \text{ with } \omega_{n,j} = r_{n,M} \|f_j\|_n$$

where $\|g\|_n^2 = n^{-1} \sum_{i=1}^n g^2(X_i)$ for the squared empirical L_2 norm of any function $g : \mathcal{X} \rightarrow \mathbb{R}$. The tuning sequence $r_{n,M} > 0$ is defined by $r_{n,M} := A\sqrt{\log(M)/n}$ for A large enough. Then we have $\hat{f}_n = f_{\hat{\lambda}}$

Define

$$M(\lambda) = \sum_{j=1}^M I_{\{\lambda_j \neq 0\}}$$

the number of non-zero coordinates of λ .

We recall the definition of weak sparsity in [BTW07]. Let $C_f > 0$ be a constant depending only on f and

$$\Lambda = \{\lambda \in \mathbb{R}^M : \|f_\lambda - f\|^2 \leq C_f r_{n,M}^2 M(\lambda)\}$$

where

$$\|g\|^2 = \int_{\mathcal{X}} g^2(x) \mu(dx)$$

If Λ is not empty, f has the weak sparsity property relative to the dictionary $\{f_1, \dots, f_M\}$.

We have then the following theorem

Theorem 16 Assume the general assumptions (A1)-(A3) and consider the notations in [BTW07]. Then, for all $\lambda \in \Lambda$,

$$\Pr(|E_{X,Y}(Y - \hat{f}_n(X))^2 - E_{X,Y}(Y - \hat{f}_{n-1}(X))^2| > 2B_1\kappa_M^{-1}r_{n,M}^2M(\lambda)) \leq 2\pi_{n-1,M}(\lambda)$$

with $\pi_{n-1,M}(\lambda)$ a small probability defined in [BTW07].

In other words, the Lasso-type algorithm is weakly stable with respect to d_e and $\|\cdot\|_1$.

Proof

According to theorem 2.1. in [BTW07], we have:

$$\Pr(E_{X,Y}|\hat{f}_n(X)) - f(X)|^2 \leq B_1\kappa_M^{-1}r_{n,M}^2M(\lambda)) \geq 1 - \pi_{n,M}(\lambda).$$

Thus, denote $\pi := \Pr(|E_{X,Y}(Y - \hat{f}_n(X))^2 - E_{X,Y}(Y - \hat{f}_{n-1}(X))^2| > 2B_1\kappa_M^{-1}r_{n,M}^2M(\lambda))$. We obtain:

$$\begin{aligned} \pi &= \Pr(|E_X(f(X) - \hat{f}_n(X))^2 - E_{X,Y}(f(X) - \hat{f}_{n-1}(X))^2| > 2B_1\kappa_M^{-1}r_{n,M}^2M(\lambda)) \\ &\leq \Pr(E_X(f(X) - \hat{f}_n(X))^2 > B_1\kappa_M^{-1}r_{n,M}^2M(\lambda)) \\ &\quad + \Pr(E_X(f(X) - \hat{f}_{n-1}(X))^2 > B_1\kappa_M^{-1}r_{n,M}^2M(\lambda)) \\ &\leq 2\pi_{n-1,M}(\lambda). \end{aligned}$$

□

As seen in the following table, we retrieve with those notations the different notions of stability introduced by [DEWA79], [KEA95] and also [BE01], [KUNY02].

stability \ distance	d_∞	d_1	d_e
Weak	weak (λ, δ) hypothesis stability [KUNY02]	weak (λ, δ) L_1 stability [KUNY02]	weak (λ, δ) error stability [KUNY02]
Strong	strong (λ, δ) hypothesis stability [KUNY02][DEWA79]	strong (λ, δ) L_1 stability [KUNY02]	strong (λ, δ) error stability [KUNY02]
Sure Stability	uniform stability [BE01]	[DEWA79]	error stability [KEA95]

To motivate this approach, we also quote a list of class of predictors satisfying the previous stability conditions.

stability distance	d_∞	d_1	d_e
Weak			Lasso
Strong	Adaboost ([KUNY02])	-ERM ([KUNY02]) - k -nearest rule	Bayesian algorithm [KEA95]
Uniform	Regularization networks		

Remark 17 We omit other weaker definition of stability such as defined in [BE01], [DEWA79], and [KUNY02]. They consider bounds on the first moment of $\mathbb{E}_{\mathcal{D}_n} d(\psi_{U_n}, \psi_n)$ instead of probability bounds. Under these assumptions, they obtain polynomial upper bounds on $\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon)$. It is would be interesting to explore the behaviour of cross-validation estimates under these hypotheses. However, this cannot be done with the techniques presented in this paper and is left to further investigation.

The main notations and definitions are summarized in the next table:

Name	Notation	Definition
Risk or generalization error	\tilde{R}_n	$E_P[L(Y, \phi(X, D_n)) \mid D_n]$
Resubstitution error	\hat{R}_n	$\frac{1}{n} \sum_{i=1}^n L(Y_i, \phi_n(X_i, D_n))$
Cross-validation error	\hat{R}_{CV}	$E_{V_n^{tr}} P_{n, V_n^{ts}} \psi_{V_n^{tr}}$

Table 2: Main notations

3 Results for risk assessment for stable algorithms

Our goal is now to derive upper bounds for the probability that the distance between the cross-validation estimator and the generalization error is greater than $\varepsilon \geq 0$: $\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon)$.

3.1 Hypotheses \mathcal{H}

Let \mathcal{D}_n be a learning set of size n . Let $V_n^{tr} \sim \mathbb{Q}$ be a training vector independent of \mathcal{D}_n such that the cross-validation is symmetric -i.e. $\Pr(V_{n,i}^{tr} = 1)$ is a constant independent of i -and the number of elements in the training set is equal to np_n . Let d be a distance among d_e, d_1, d_∞ . At last, we suppose that the loss function L is bounded by 1. We derive the following general results that stands for general cross-validation procedures and stable algorithms.

3.2 Strong stability

We state two results according to the class of stability. We will use the definition of strong difference bounded introduced by [KUT02] and a corollary of his main theorem inspired by [McD89].

Definition 18 (Kutin[KUT02]) Let $\Omega_1, \dots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^n \Omega_k$ and let X a random variable on Ω . We say that X is strongly difference bounded by (b, c, δ) if the following holds: there is a "bad" subset $B \subset \Omega$, where $\delta = \mathbb{P}(B)$. If $\omega, \omega' \in \Omega$ differ only in k -th coordinate, and $\omega \notin B$, then

$$|X(\omega) - X(\omega')| \leq c.$$

Furthermore, for any $\omega, \omega' \in \Omega$,

$$|X(\omega) - X(\omega')| \leq b.$$

We will need the following theorem. It says in substance that a strongly difference bounded function of independent variables is closed to its expectation with high probability.

Theorem 19 (Kutin[KUT02]) Let $\Omega_1, \dots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^n \Omega_k$ and let X a random variable on Ω , which is strongly difference bounded by (b, c, δ) . Assume $b \geq c \geq 0$ and $\alpha' > 0$. Let $\mu = \mathbb{E}(X)$. Then, for any $\tau > 0, \alpha' > 0$,

$$\Pr(X - \mu \geq \tau) \leq 2(\exp(-\frac{\tau^2}{8n(c + b\alpha')^2}) + \frac{n}{\alpha'}\delta).$$

We are now in position to derive

Theorem 20 (Cross-validation strong stability) Suppose that \mathcal{H} holds. Let Ψ a machine learning which is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable. Then, for all $\varepsilon \geq 0$,

$$\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}.$$

Furthermore, if d is the uniform distance d_∞ , then we have for all $\varepsilon \geq 0$:

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n(1-p_n)} + \lambda(2p_n)^\alpha) \leq 4(\exp(-\frac{\varepsilon^2}{8n(5\lambda(2p_n)^\alpha + \alpha')^2}) + \frac{n}{\alpha'}\delta'_{n,p_n}),$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n+1,1/(n+1)}$.
Thus, if we choose $\alpha = 5\lambda(2np_n)^\alpha$,

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) \leq 4(\exp(-\frac{\varepsilon^2}{8(10\lambda)^2n(2p_n)^{2\alpha}}) + \frac{n}{5\lambda(2p_n)^\alpha}\delta'_{n,p_n}).$$

Proof

1. For the general case, denote B the bad subset, i.e. $B = \{\sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{\|\mathbb{P}_{n,U_n} - \mathbb{P}_n\|_\alpha} \geq \lambda\}$. Since Ψ is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d, \mathbb{Q})$ stable, we have $\Pr(B) \leq \delta_{n,p_n}$. It is sufficient to split $|\widehat{R}_{CV} - \widetilde{R}_n|$ according to a benchmark, namely $\overline{R}_{n(1-p_n)} := \mathbb{E}_{V_n^{tr}} \mathbb{P} \psi_{V_n^{tr}}$. Thus, we get

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq \Pr(|\widehat{R}_{CV} - \overline{R}_{n(1-p_n)}| \geq \varepsilon) + \Pr(|\overline{R}_{n(1-p_n)} - \widetilde{R}_n| \geq \lambda(2p_n)^\alpha)$$

The first term can be bounded by conditional Hoeffding inequality (see chapter 1). Thus, we obtain

$$\Pr(|\widehat{R}_{CV} - \overline{R}_{n(1-p_n)}| \geq \varepsilon) \leq 2 \exp(-2np_n\varepsilon^2).$$

For the second term, notice that:

$$|\overline{R}_{n(1-p_n)} - \widetilde{R}_n| = |\mathbb{E}_{V_n^{tr}} \mathbb{P} \psi_{V_n^{tr}} - \mathbb{P} \psi_n| \leq \mathbb{E}_{V_n^{tr}} |\mathbb{P} \psi_{V_n^{tr}} - \mathbb{P} \psi_n|.$$

Recall that for any $d \in \{d_e, d_1, d_\infty\}$, we have $|\mathbb{P} \psi_{V_n^{tr}} - \mathbb{P} \psi_n| \leq d(\psi_{V_n^{tr}}, \psi_n)$ and $\|\mathbb{P}_{n,V_n^{tr}} - \mathbb{P}_n\|_\alpha = (2p_n)^\alpha$.

Thus, since Ψ is strong $(\lambda, (\delta_n)_n)$ stable, we have

$$\begin{aligned} \Pr(|\overline{R}_{n(1-p_n)} - \widetilde{R}_n| \geq \lambda(2p_n)^\alpha) &\leq \Pr(\sup_{V_n^{tr} \in \text{support}(\mathbb{Q})} d(\psi_{V_n^{tr}}, \psi_n) / \|\mathbb{P}_{n,V_n^{tr}} - \mathbb{P}_n\|_\alpha \geq \lambda) \\ &= \Pr(B) \leq \delta_{n,p_n} \end{aligned}$$

2. In the particular case, when $d = d_\infty$, the most stable notion of stability, we can obtain a stronger result. For this, we recall two very useful results.

We proceed in three steps as in [BE02],[KUNIY02] by using a bounded difference inequality

- first, we show that the expectation of $\widehat{R}_{CV} - \widetilde{R}_n$ is small,
- secondly, we show that the function $\widehat{R}_{CV} - \widetilde{R}_n$ seen as a function f of Z_1, Z_2, \dots, Z_n is strongly difference bounded, i.e.: with high probability, there exists constants c_1, \dots, c_n such that we have for all i , for all $z \in \mathcal{Z}$,

$$|f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, Z_{i-1}, z, Z_{i+1}, \dots, Z_n)| \leq c_i,$$

- use theorem 19 with the first two points,
- at least, use arguments of symmetry to conclude.

1. The expectation of $\widehat{R}_{CV} - \widetilde{R}_n$ is small

Let us denote $\mathbf{v}_n^{tr}, \mathbf{v}_n^{ts}$ fixed training and test vectors.

$$\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) = \mathbb{P}^{\otimes n}(\mathbb{E}_{V_n^{tr} \mathbb{P}_{n, V_n^{ts}}} \psi_{V_n^{tr}} - \mathbb{P} \psi_n) = \mathbb{P}^{\otimes n} \mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n)$$

since

$$\mathbb{P}^{\otimes n} \mathbb{E}_{V_n^{tr} \mathbb{P}_{n, V_n^{ts}}} \psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}} \mathbb{P}^{\otimes n} \mathbb{P}_{n, V_n^{ts}} \psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}} \mathbb{P}^{\otimes n} \mathbb{P} \psi_{V_n^{tr}} = \mathbb{P}^{\otimes n} \mathbb{P} \psi_{\mathbf{v}_n^{tr}}$$

where the first equality comes from the linearity of expectation, the second from the fact that $\mathbb{P}_{n, V_n^{tr}}$ are independent of $\mathbb{P}_{n, V_n^{ts}}$, and the third from the i.i.d. nature of $(Z_i)_i$.

Recall that $\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$ where d stands indifferently for d_1, d_e or d_∞ . Thus, $\mathbb{P}^{\otimes n} \mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq \mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$. By conditioning according to the small values of $d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$, we obtain

$$\begin{aligned} \mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) &= \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) | B) \mathbb{P}^{\otimes n}(B) + \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) | B^c) (1 - \mathbb{P}^{\otimes n}(B)) \\ &\leq 1 \times \delta_{n, p_n} + \lambda \mathbb{P}^{\otimes n} \|\mathbb{P}_{n, \mathbf{v}_n^{tr}} - \mathbb{P}_n\|_\alpha \times (1 - \delta_{n, p_n}) = \delta_{n, p_n} + \lambda (2p_n)^\alpha (1 - \delta_{n, p_n}) \end{aligned}$$

Eventually, we get $\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) \leq \delta_{n, p_n} + \lambda (2p_n)^\alpha$.

2. $\widehat{R}_{CV} - \widetilde{R}_n$ is difference bounded with high probability

Denote $f(Z_1, Z_2, \dots, Z_n) := \widehat{R}_{CV} - \widetilde{R}_n$. Let $z \in \mathcal{Z}$. Let $\mathcal{D}_{n+1} = \mathcal{D}_{n+1} \cup \{z\}$. Now denote $B = B_1 \cup B_2$ where

$$B_1 = \left\{ \sup_{U_n \in \text{support}(\mathbb{Q})} \frac{d(\psi_{U_n}, \psi_n)}{\|\mathbb{P}_{n, U_n} - \mathbb{P}_n\|_\alpha} \geq \lambda \right\}$$

and

$$B_2 = \left\{ \sup_{1 \leq i \leq n+1} \frac{d(\psi_{e_{n+1}^i}, \psi_{n+1})}{\|\mathbb{P}_{n+1, e_{n+1}^i} - \mathbb{P}_{n+1}\|_\alpha} \geq \lambda \right\}$$

with e_{n+1}^i the binary of size $n+1$ equal to 0 everywhere except on the i -th coordinate $e_{n+1, k}^i := 1_{(k=i)}$ for $1 \leq k \leq n+1$. Under our assumptions, we have

$$\Pr(B) \leq \delta_{n, p_n} + (n+1) \delta_{n+1, 1/n+1}$$

.

We want to show that with high probability there exist constants c_i such that for all $i \in \{1, \dots, n\}$, for all $z \in \mathcal{Z}$, $|f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, Z_{i-1}, z, Z_{i+1}, \dots, Z_n)| \leq c_i$.

Notice that

$$\begin{aligned} |f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, z, \dots, Z_n)| &= |(\mathbb{E}_{V_n^{tr} \mathbb{P}_{n, V_n^{ts}}} \psi_{V_n^{tr}} - \mathbb{P} \psi_{e_{n+1}^{n+1}}) \\ &\quad - (\mathbb{E}_{V_n^{tr} \mathbb{P}_{n, V_n^{ts}}} \psi_{V_n^{tr}} - \mathbb{P} \psi_{e_{n+1}^i})| \\ &\leq |\mathbb{E}_{V_n^{tr} \mathbb{P}_{n, V_n^{ts}}} \psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr} \mathbb{P}_{n, V_n^{ts}}} \psi_{V_n^{tr}}'| \\ &\quad + |\mathbb{P} \psi_{e_{n+1}^{n+1}} - \mathbb{P} \psi_{e_{n+1}^i}|. \end{aligned}$$

with $\mathbb{P}'_{n,V_n^{tr}}$ the weighted empirical measure on the sample

$$\mathcal{E}_n = \{Z_1, \dots, Z_{i-1}, z, Z_{i+1}, \dots, Z_n\}$$

and $\psi'_{V_n^{tr}}$ the predictor trained on $\mathcal{E}_{V_n^{tr}}$.

So, first, let us bound the second term, recall that

$$|\mathbb{P}(\psi_{e_{n+1}^{n+1}} - \psi_{e_{n+1}^i})| \leq d(\psi_{e_{n+1}^{n+1}}, \psi_{e_{n+1}^i}) \leq d(\psi_{e_{n+1}^{n+1}}, \psi_{n+1}) + d(\psi_{n+1}, \psi_{e_{n+1}^i})$$

with ψ_{n+1} trained on $\mathcal{D}_{n+1} = \{Z_1, \dots, Z_{i-1}, Z_i, Z_{i+1}, \dots, Z_n, z\}$. Thus, we have on B^C , $|\mathbb{P}\psi_{e_{n+1}^{n+1}} - \mathbb{P}\psi_{e_{n+1}^i}| \leq 2(\frac{2\lambda}{n+1})^\alpha$.

To upper bound the first term, notice that

$$\begin{aligned} |\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n,V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}} \mathbb{P}'_{n,V_n^{ts}} \psi'_{V_n^{tr}}| &= |\mathbb{E}_{V_n^{tr}} (\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}}) | V_{n,i}^{tr} = 1) \times (1 - p_n) \\ &\quad + \mathbb{E}_{V_n^{tr}} ((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}}) \psi_{V_n^{tr}} | V_{n,i}^{ts} = 1) \times p_n|. \end{aligned}$$

We always have for any ψ , $|(\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}}) \psi| \leq 1/n p_n$ thus

$$|\mathbb{E}_{V_n^{tr}} ((\mathbb{P}_{n,V_n^{ts}} - \mathbb{P}'_{n,V_n^{ts}}) \psi_{V_n^{tr}} | V_{n,i}^{ts} = 1) \times p_n| \leq 1/n$$

Until now, the previous lines hold independently of $d \in \{d_e, d_1, d_\infty\}$. We still have to bound $|\mathbb{E}_{V_n^{tr}} (\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}}) | V_{n,i}^{tr} = 1)|$. In the particular case of the most stable kind of stability (i.e. when $d = d_\infty$), we have

$$|\mathbb{E}_{V_n^{tr}} (\mathbb{P}_{n,V_n^{ts}}(\psi_{V_n^{tr}} - \psi'_{V_n^{tr}}) | V_{n,i}^{tr} = 1)| \leq \mathbb{E}_{V_n^{tr}} (d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}) | V_n^{tr} = 1).$$

On B^C , we get $d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}) \leq d_\infty(\psi_{V_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi'_{V_n^{tr}}) \leq 2(2\lambda p_n)^\alpha$.

Thus, on B^C , we have

$$\mathbb{E}_{V_n^{tr}} (d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}) | V_{n,i}^{tr} = 1) \leq 2(2\lambda p_n)^\alpha.$$

Putting all together, with probability at least $1 - \delta'_{n,p_n}$,

$$\sup_{1 \leq i \leq n, z \in \mathcal{Z}} |f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, z, \dots, Z_n)| \leq 5(2\lambda p_n)^\alpha.$$

3. $\widehat{R}_{CV} - \widetilde{R}_n$ is closed to zero with high probability

Applying theorem 19, we obtain that for all $\varepsilon \geq 0$

$$\begin{aligned} \Pr(\widehat{R}_{CV} - \widetilde{R}_n \geq \varepsilon + \delta + \lambda(2p_n)^\alpha) &\leq \Pr(\widehat{R}_{CV} - \widetilde{R}_n - \mathbb{E}_{\mathcal{D}_n}(\widehat{R}_{CV} - \widetilde{R}_n) \geq \varepsilon) \\ &\leq 2(\exp(-\frac{\varepsilon^2}{8n(5(2\lambda p_n)^\alpha + \alpha')^2}) + \frac{n}{\alpha'} \delta') \\ &\leq 2(\exp(-\frac{\varepsilon^2}{8(10\lambda)^2 n(2\lambda p_n)^{2\alpha}}) + \frac{n}{5(2\lambda p_n)^\alpha} \delta') \\ &\text{by taking } \alpha' = 5(2\lambda p_n)^\alpha. \end{aligned}$$

By symmetry, we also have $\Pr(\widehat{R}_{CV} - \widetilde{R}_n \leq -(\varepsilon + \delta_{n,p_n} + 2\lambda p_n)) \leq 2(\exp(-\frac{\tau^2}{8(10\lambda)^2 n(2\lambda p_n)^{2\alpha}}) + \frac{n}{5(2\lambda p_n)^\alpha} \delta'_{n,p_n})$ which allows to conclude. \square

Theorem 21 (Strong stability) *Suppose that \mathcal{H} holds. Let Ψ be a machine learning which is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable. Then, for all $\varepsilon \geq 0$, we get*

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq 2\exp(-2np_n\varepsilon^2) + \kappa(n)\delta_n,$$

where $\kappa(n)$ is the number of training vectors in the cross-validation.

Furthermore, if the distance d is the uniform distance d_∞ , then we have for any $\varepsilon \geq 0$:

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) \leq 4(\exp(-\frac{\varepsilon^2}{8n(5(2\lambda p_n)^\alpha + \alpha')^2}) + \frac{n}{\alpha'} \kappa(n) \delta'_{n,p_n}),$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n,1/n}$. Thus, if we take $\alpha = 5(2\lambda p_n)^\alpha$, we get

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) \leq 4(\exp(-\frac{\varepsilon^2}{8(10\lambda)^2 n(2\lambda p_n)^{2\alpha}}) + \frac{n}{5(2\lambda p_n)^\alpha} \kappa(n) \delta'_{n,p_n}).$$

Proof

For the first inequality, it is sufficient to use remarks 8.

For the second one, we can follow the previous proof, using remarks 8 and noticing that if we denote $B_{\mathbf{v}_n^{tr}} := \{d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) \geq \lambda \|\mathbb{P}_{n,\mathbf{v}_n^{tr}} - \mathbb{P}_n\|\}$, then, we have

$$\begin{aligned} \mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) &= \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) | B_{\mathbf{v}_n^{tr}}) \mathbb{P}^{\otimes n}(B_{\mathbf{v}_n^{tr}}) + \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) | B_{\mathbf{v}_n^{tr}}^c) (1 - \mathbb{P}^{\otimes n}(B_{\mathbf{v}_n^{tr}})) \\ &\leq 1 \times \delta_{n,p_n} + \lambda \mathbb{P}^{\otimes n} \|\mathbb{P}_{n,\mathbf{v}_n^{tr}} - \mathbb{P}_n\|_\alpha \times (1 - \delta_{n,p_n}) = \delta_{n,p_n} + \lambda(2p_n)^\alpha (1 - \delta_{n,p_n}). \end{aligned}$$

Eventually, we get $\mathbb{P}^{\otimes n}(\widehat{R}_{CV} - \widetilde{R}_n) \leq \delta_{n,p_n} + \lambda(2p_n)^\alpha$. \square

Now, we derive results for the hold-out cross-validation which does not make a symmetrical use of the dataset. We obtain

Theorem 22 (Strong stability and hold-out) *Let Ψ be a machine learning which is strong $(\lambda, (\delta_{n,p_n})_{n,p_n}, \delta)$ stable. Then the hold-out (or split sample) cross-validation satisfies for all $\varepsilon \geq 0$,*

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq 2\exp(-2np_n\varepsilon^2) + \delta_{n,p_n}.$$

Furthermore, if the distance is the uniform distance d_∞ , then we have

$$\begin{aligned} \Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) &\leq 4(\exp(-\frac{\varepsilon^2}{8(4\lambda(2p_n)^\alpha + 1/np_n)^2}) \\ &\quad + \frac{n^2}{4\lambda(2p_n)^\alpha + 1/np_n} \delta'_{n,p_n}), \end{aligned}$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + n\delta_{n,1/n}$

Proof

For the first inequality, it is enough to use remarks 8.

For the second one, we start as previously. First, we bound in the same way the expectation.

Secondly, we show that $\widehat{R}_{CV} - \widetilde{R}_n$ is difference-bounded with high probability.

Denote $f(Z_1, Z_2, \dots, Z_n) := \widehat{R}_{CV} - \widetilde{R}_n$. Let $z \in \mathcal{Z}$. Now denote as previously $B := B_1 \cup B_2$ with

$$B_1 = \left\{ \frac{d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)}{\|\mathbb{P}_{n, \mathbf{v}_n^{tr}} - \mathbb{P}_n\|_\alpha} \geq \lambda \right\} \text{ and } B_2 = \left\{ \sup_i \frac{d(\psi_{e_{n+1}^i}, \psi_{n+1})}{\|\mathbb{P}_{n+1, e_{n+1}^i} - \mathbb{P}_{n+1}\|_\alpha} \geq \lambda \right\}. \text{ Eventually, we have } \Pr(B) \leq \delta_{n, 1-p_n} + n\delta_{n, 1/n}$$

We want to show that with high probability there exists constants c_i such that for all i , for all $z \in \mathcal{Z}$, $|f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, z, \dots, Z_n)| \leq c_i$. Since $V_n^{tr} = \mathbf{v}_n^{ts}$ fixed vector in the case of hold-out, notice that:

$$\begin{aligned} |f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, z, \dots, Z_n)| &= |\mathbb{P}_{n, \mathbf{v}_n^{tr}} \psi_{\mathbf{v}_n^{tr}} - \mathbb{P}_{e_{n+1}^{n+1}} - (\mathbb{P}'_{n, \mathbf{v}_n^{ts}} \psi'_{\mathbf{v}_n^{tr}} - \mathbb{P} \psi_{e_{n+1}^i})| \\ &\leq |\mathbb{P}_{n, \mathbf{v}_n^{tr}} \psi_{\mathbf{v}_n^{tr}} - \mathbb{E}_{\mathbf{v}_n^{tr}} \mathbb{P}'_{n, \mathbf{v}_n^{ts}} \psi'_{\mathbf{v}_n^{tr}}| \\ &\quad + |\mathbb{P} \psi_{e_{n+1}^{n+1}} - \mathbb{P} \psi_{e_{n+1}^i}|, \end{aligned}$$

with $\mathbb{P}'_{n, \mathbf{v}_n^{tr}}$ the weighted empirical measures of the sample $\mathcal{E}_n = \{Z_1, \dots, Z_{i-1}, z, Z_{i+1}, \dots, Z_n\}$ and $\psi'_{\mathbf{v}_n^{tr}}$ the predictor trained on $\mathcal{E}_{\mathbf{v}_n^{tr}}$.

So, first, let us bound the second term, recall that:

$$|\mathbb{P}(\psi_{e_{n+1}^{n+1}} - \psi_{e_{n+1}^i})| \leq d(\psi_{e_{n+1}^{n+1}}, \psi_{e_{n+1}^i}) \leq d(\psi_{e_{n+1}^{n+1}}, \psi_{n+1}) + d(\psi_{n+1}, \psi_{e_{n+1}^i})$$

Thus, on B^c , $|\mathbb{P} \psi_{e_{n+1}^{n+1}} - \mathbb{P} \psi_{e_{n+1}^i}| \leq 2\lambda(\frac{2}{n+1})^\alpha$.

To upper bound the first term, notice that:

$$|\mathbb{P}_{n, \mathbf{v}_n^{tr}} \psi_{\mathbf{v}_n^{tr}} - \mathbb{P}'_{n, \mathbf{v}_n^{ts}} \psi'_{\mathbf{v}_n^{tr}}| = |\mathbb{P}_{n, \mathbf{v}_n^{tr}} (\psi_{\mathbf{v}_n^{tr}} - \psi'_{\mathbf{v}_n^{tr}}) 1_{\{\mathbf{v}_{n,i}^{tr}=1\}} + (\mathbb{P}_{n, \mathbf{v}_n^{tr}} - \mathbb{P}'_{n, \mathbf{v}_n^{ts}}) \psi_{\mathbf{v}_n^{tr}} 1_{\{\mathbf{v}_{n,i}^{tr}=1\}}|$$

We always have for any ψ , $|\mathbb{P}_{n, \mathbf{v}_n^{tr}} - \mathbb{P}'_{n, \mathbf{v}_n^{ts}}| \leq 1/n p_n$ thus

$$|(\mathbb{P}_{n, \mathbf{v}_n^{tr}} - \mathbb{P}'_{n, \mathbf{v}_n^{ts}}) \psi_{\mathbf{v}_n^{tr}} 1_{\{\mathbf{v}_{n,i}^{tr}=1\}}| \leq 1/n p_n$$

We still have to bound $|\mathbb{P}_{n, \mathbf{v}_n^{tr}} (\psi_{\mathbf{v}_n^{tr}} - \psi'_{\mathbf{v}_n^{tr}}) 1_{\{\mathbf{v}_{n,i}^{tr}=1\}}|$. As in the previous proof, we have when $d = d_\infty$,

$$|\mathbb{P}_{n, \mathbf{v}_n^{tr}} (\psi_{\mathbf{v}_n^{tr}} - \psi'_{\mathbf{v}_n^{tr}}) 1_{\{\mathbf{v}_{n,i}^{tr}=1\}}| \leq d_\infty (\psi_{\mathbf{v}_n^{tr}}, \psi'_{\mathbf{v}_n^{tr}}) 1_{\{\mathbf{v}_{n,i}^{tr}=1\}}$$

On B^c , $d_\infty(\psi_{\mathbf{v}_n^{tr}}, \psi'_{\mathbf{v}_n^{tr}}) \leq d_\infty(\psi_{\mathbf{v}_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi'_{\mathbf{v}_n^{tr}}) \leq 2\lambda(2p_n)^\alpha$. Thus, on B^c we get $d_\infty(\psi_{\mathbf{v}_n^{tr}}, \psi'_{\mathbf{v}_n^{tr}}) 1_{\{\mathbf{v}_{n,i}^{tr}=1\}} \leq 2\lambda(2p_n)^\alpha$.

Putting all together, with probability at least $1 - \delta'_{n, p_n}$,

$$\begin{aligned} \sup_{i, z} |f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, z, \dots, Z_n)| &\leq 2\lambda\left(\frac{2}{n+1}\right)^\alpha + \max((np_n)^{-1}, 2\lambda(2p_n)^\alpha) \\ &\leq 4\lambda(2p_n)^\alpha + (np_n)^{-1} \end{aligned}$$

To conclude, apply again theorem 19.

□

3.3 Weak stability

We now derive results that stands for general cross-validation procedures and weakly stable predictors. We recall here the interest of the notion of weak stability. For some class of machine learning, the notion of strong stability may be too demanding. That is why weak stability is introduced. As a motivation, algorithms such as Adaboost satisfies the following definition of weak stability.

We will use the definition of weak difference bounded introduced by [KUT02] and a corollary of his main theorem.

Definition 23 (Kutin[KUT02]) Let $\Omega_1, \dots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^n \Omega_k$ and let X a random variable on Ω . We say that X is weakly difference bounded by (b, c, δ) if the following holds: for any k ,

$$\forall^\delta(\omega, v) \in \Omega \times \Omega_k, \mathbb{P}(|X(\omega) - X(\omega')|) \leq c$$

where $\omega'_k = v$ and $\omega'_i = \omega_i$ for $i \neq k$. and the notation $\forall^\delta \omega, \Phi(\omega)$ means " $\Phi(\omega)$ holds for all but but a δ fraction of Ω "

$$|X(\omega) - X(\omega')| \leq c$$

Furthermore, for any $\omega, \omega' \in \Omega$, differing only one coordinate:

$$|X(\omega) - X(\omega')| \leq b$$

We will need the following theorem. It says in substance that a weakly difference bounded function of independent variables is closed to its expectation with probability.

Theorem 24 (Kutin[KUT02]) Let $\Omega_1, \dots, \Omega_n$ be probability spaces. Let $\Omega = \prod_{k=1}^n \Omega_k$ and let X a random variable on Ω . which is weakly difference bounded by (b, c, δ) . Assume $b \geq c \geq 0$ and $\alpha > 0$. Let $\mu = \mathbb{E}(X)$. Then, for any $\varepsilon > 0$

$$\Pr(|X - \mu| \geq \varepsilon) \leq 2 \exp\left(-\frac{\varepsilon^2}{10nc^2(1 + \frac{2\varepsilon}{15nc})^2}\right) + \frac{2nb\delta^{1/2}}{c} \exp\left(\frac{\varepsilon b}{4nc^2}\right) + 2n\delta^{1/2}.$$

Theorem 25 (Cross-validation Weak stability) Suppose that \mathcal{H} holds. Let Ψ be a machine learning which is weak $(\lambda, (\delta_n)_n, d, \mathbb{Q})$ stable with respect to the distance d . Then, for all $\varepsilon \geq 0$,

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq 2 \exp(-2np_n\varepsilon^2) + \delta_{n,p_n}$$

Furthermore, if the distance is the uniform distance d_∞ , we have for all $\varepsilon \geq 0$:

$$\begin{aligned} \Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) &\leq 4 \left(\exp\left(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^\alpha)^2(1 + \frac{2\varepsilon}{15n(5\lambda(2p_n)^\alpha)})^2}\right) \right. \\ &\quad \left. + \frac{2n\delta'_{n,p_n}}{5\lambda(2p_n)^\alpha} \exp\left(\frac{\varepsilon n}{4n(5\lambda(2p_n)^\alpha)^2}\right) + n\delta'_{n,p_n} \right), \end{aligned}$$

with $\delta'_{n,p_n} = 2\delta_{n,1/n} + \delta_{n,p_n}$

Proof

In the following, denote B the bad subset, i.e. $B = \cup_{v_n^{tr}} B_{v_n^{tr}}$ with $B_{v_n^{tr}} = \{d(\psi_{\mathbb{P}_{n,v_n^{tr}}}, \psi_{\mathbb{P}_n}) \geq \lambda|\mathbb{P}_{n,v_n^{tr}} - \mathbb{P}_n|\}$. Since Ψ is strong $(\lambda, (\delta_n)_{n,p_n}, d, \mathbb{Q})$ stable, we have $\mathbb{P}(B) \leq \delta_{n,p_n}$.

1. For the general case, it is again sufficient to split $|\hat{R}_{CV} - \tilde{R}_n|$ according to the same benchmark, namely $\bar{R}_{n(1-p)} = \mathbb{E}_{V_n^{tr}} \mathbb{P} \psi_{V_n^{tr}}$.

Thus,

$$\Pr(|\hat{R}_{CV} - \tilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq \Pr(|\hat{R}_{CV} - \bar{R}_{n(1-p)}| \geq \varepsilon) + \Pr(|\bar{R}_{n(1-p)} - \tilde{R}_n| \geq \lambda(2p_n)^\alpha)$$

The first term can be bounded as previously by $2 \exp(-2np_n\varepsilon^2)$.

For the second term, notice that $|\bar{R}_{n(1-p)} - \tilde{R}_n| = |\mathbb{E}_{V_n^{tr}} \mathbb{P} \psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}} \mathbb{P} \psi_n| \leq \mathbb{E}_{V_n^{tr}} |\mathbb{P} \psi_{V_n^{tr}} - \mathbb{P} \psi_n|$. Recall that $|\mathbb{P} \psi_{V_n^{tr}} - \mathbb{P} \psi_n| \leq d(\psi_{V_n^{tr}}, \psi_n)$ and $\|\mathbb{P}_{n, V_n^{tr}} - \mathbb{P}_n\|_\alpha^\alpha = \lambda(2p_n)^\alpha$. Thus, since Ψ is weak $(\lambda, (\delta_{n, p_n})_{n, p_n}, d)$ stable, we have

$$\begin{aligned} \Pr(|\hat{R}_{n(1-p_n)} - \tilde{R}_n| \geq \lambda(2p_n)^\alpha) &\leq \Pr(\mathbb{E}_{v_n^{tr}} d(\psi_{v_n^{tr}}, \psi_n) \geq \lambda(2p_n)^\alpha) \\ &\leq \Pr(\cup_{v_n^{tr}} \{d(\psi_{v_n^{tr}}, \psi_n) \geq \lambda \|\mathbb{P}_{n, v_n^{tr}} - \mathbb{P}_n\|_\alpha\}) \\ &= \Pr(\cup_{v_n^{tr}} B_{v_n^{tr}}) \leq \kappa(n) \delta_{n, p_n}. \end{aligned}$$

2. In the particular case, when $d = d_\infty$, we can also obtain a stronger result.

We proceed in three steps as in [BE02], [KUNIY02] by using a bounded difference inequality:

1. first, we show that the expectation of $\hat{R}_{CV} - \tilde{R}_n$ is small of the same order as for the strong stability.
2. secondly, we show that the function $\hat{R}_{CV} - \tilde{R}_n$ seen as a function f of Z_1, Z_2, \dots, Z_n is weakly difference bounded, i.e. there exists constants c_1, \dots, c_n such that for all i , if $Z_1, \dots, Z_i, \dots, Z_n, Z_{i'}$ i.i.d. random variables, we have with high probability

$$|f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, Z_{i'}, \dots, Z_n)| \leq c_i.$$

3. finally, we use theorem 24 with the first two points to conclude.

1. The expectation of $\hat{R}_{CV} - \tilde{R}_n$ is small

As previously, denote $\mathbf{v}_n^{tr}, \mathbf{v}_n^{ts}$ fixed vectors. We still have

$$\mathbb{P}^{\otimes n}(\hat{R}_{CV} - \tilde{R}_n) = \mathbb{P}^{\otimes n}(\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n, V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{P} \psi_n) = \mathbb{P}^{\otimes n} \mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n).$$

since $\mathbb{P}^{\otimes n} \mathbb{E}_{V_n^{tr}} \mathbb{P}_{n, V_n^{ts}} \psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}} \mathbb{P}^{\otimes n} \mathbb{P}_{n, V_n^{ts}} \psi_{V_n^{tr}} = \mathbb{E}_{V_n^{tr}} \mathbb{P}^{\otimes n} \mathbb{P} \psi_{V_n^{tr}} = \mathbb{P}^{\otimes n} \mathbb{P} \psi_{\mathbf{v}_n^{tr}}$ where the first equality comes from the linearity of expectation, the second from the fact that $\mathbb{P}_{n, V_n^{tr}}$ are independent of $\mathbb{P}_{n, V_n^{ts}}$, and the third one from the *i.i.d.* nature of $(Z_i)_i$.

Recall that $\mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$ where d stands indifferently for d_1, d_e or d_∞ . Thus, $\mathbb{P}^{\otimes n} \mathbb{P}(\psi_{\mathbf{v}_n^{tr}} - \psi_n) \leq \mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$. By conditioning according to the small values of $d(\psi_{\mathbf{v}_n^{tr}}, \psi_n)$, we obtain

$$\begin{aligned} \mathbb{P}^{\otimes n} d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) &= \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) | B_{\mathbf{v}_n^{tr}}) \mathbb{P}^{\otimes n}(B_{\mathbf{v}_n^{tr}}) \\ &\quad + \mathbb{P}^{\otimes n}(d(\psi_{\mathbf{v}_n^{tr}}, \psi_n) | B_{\mathbf{v}_n^{tr}}^c) (1 - \mathbb{P}^{\otimes n}(B_{\mathbf{v}_n^{tr}})) \\ &\leq 1 \times \delta_{n, p_n} + \lambda \mathbb{P}^{\otimes n} \|\mathbb{P}_{n, \mathbf{v}_n^{tr}} - \mathbb{P}_n\|_\alpha \times (1 - \delta_{n, p_n}) \leq \delta_{n, p_n} + \lambda(2p_n)^\alpha. \end{aligned}$$

Eventually, we still have $\mathbb{P}^{\otimes n}(\hat{R}_{CV} - \tilde{R}_n) \leq \delta_{n, p_n} + \lambda(2p_n)^\alpha$.

2. $\widehat{R}_{CV} - \widetilde{R}_n$ is difference bounded with high probability

Denote $f(Z_1, Z_2, \dots, Z_n) := \widehat{R}_{CV} - \widetilde{R}_n$.

We want to show that for all i , there exists constant c_i such

$$|f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, Z'_i, \dots, Z_n)| \leq c_i$$

with high probability where $Z_1, \dots, Z_i, \dots, Z_n, Z'_i$ are i.i.d. variables. Denote

$$B_i = \left\{ \frac{d(\psi_{e_{n+1}^i}, \psi_{n+1})}{\|\mathbb{P}_{n+1, e_{n+1}^i} - \mathbb{P}_{n+1}\|_\alpha} \geq \lambda \right\}.$$

We proceed as previously where $(\cup_{v_n^{tr}} B_{v_n^{tr}}) \cup B_i \cup B_{n+1}$ will play the role of B .

$$\begin{aligned} |f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, Z'_i, \dots, Z_n)| &= |(\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n, V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{P} \psi_n) - \\ &\quad (\mathbb{E}_{V_n^{tr}} \mathbb{P}'_{n, V_n^{ts}} \psi'_{V_n^{tr}} - \mathbb{P} \psi'_n)| \\ &\leq |\mathbb{E}_{V_n^{tr}} \mathbb{P}_{n, V_n^{ts}} \psi_{V_n^{tr}} - \mathbb{E}_{V_n^{tr}} \mathbb{P}'_{n, V_n^{ts}} \psi'_{V_n^{tr}}| \\ &\quad + |\mathbb{P} \psi_n - \mathbb{P} \psi'_n|, \end{aligned}$$

with $\mathbb{P}'_n, \mathbb{P}'_{n, V_n^{ts}}$ the weighted empirical measures of the sample

$$\mathcal{D}'_n = \{Z_1, \dots, Z'_i, \dots, Z_n\}$$

and ψ'_n the predictor built on \mathcal{D}'_n .

So, first, let us bound the second term, recall that: $|\mathbb{P}(\psi_n - \psi'_n)| \leq d(\psi_n, \psi'_n) \leq d(\psi_n, \psi_{n+1}) + d(\psi_{n+1}, \psi'_n)$. with ψ_{n+1} the predictor trained on the sample $\mathcal{D}_{n+1} = \{Z_1, \dots, Z_i, \dots, Z_n, Z'_i\}$. Thus, on B^c , we have $|\mathbb{P} \psi_n - \mathbb{P} \psi'_n| \leq 2\lambda(2/n)^\alpha$.

To upper bound the first term, notice that

$$\begin{aligned} |E_{V_n^{tr}} P_{n, V_n^{ts}} \psi_{V_n^{tr}} - E_{V_n^{tr}} P'_{n, V_n^{ts}} \psi'_{V_n^{tr}}| &= |E_{V_n^{tr}} (P_{n, V_n^{ts}} (\psi_{V_n^{tr}} - \psi'_{V_n^{tr}}) | V_{n,i}^{tr} = 1) \times (1 - p_n) \\ &\quad + E_{V_n^{tr}} ((P_{n, V_n^{ts}} - P'_{n, V_n^{ts}}) \psi_{V_n^{tr}} | V_{n,i}^{tr} = 1) \times p_n|. \end{aligned}$$

We always have for all ψ , $|(\mathbb{P}_{n, V_n^{ts}} - \mathbb{P}'_{n, V_n^{ts}}) \psi| \leq 1/n p_n$ thus we get

$$|\mathbb{E}_{V_n^{tr}} ((\mathbb{P}_{n, V_n^{ts}} - \mathbb{P}'_{n, V_n^{ts}}) \psi_{V_n^{tr}} | V_n^{ts} = 1) \times p_n| \leq 1/n$$

We still have to bound

$$|\mathbb{E}_{V_n^{tr}} (\mathbb{P}_{n, V_n^{ts}} (\psi_{V_n^{tr}} - \psi'_{V_n^{tr}}) | V_{n,i}^{tr} = 1)| \leq \mathbb{E}_{V_n^{tr}} (d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}) | V_{n,i}^{tr} = 1)$$

On $B_{v_n^{tr}}^c$, $d_\infty(\psi_{v_n^{tr}}, \psi'_{v_n^{tr}}) \leq d_\infty(\psi_{v_n^{tr}}, \psi_{n+1}) + d_\infty(\psi_{n+1}, \psi'_{v_n^{tr}}) \leq 2\lambda(2p_n)^\alpha$.

Thus, we get $\mathbb{E}_{V_n^{tr}} (d_\infty(\psi_{V_n^{tr}}, \psi'_{V_n^{tr}}) | V_n^{tr} = 1) \leq 2\lambda(2p_n)^\alpha$ on $(\cup_{v_n^{tr}} B_{v_n^{tr}})^c$.

Putting all together, with probability at least $1 - 2\delta_{n+1, 1/(n+1)} - \delta_{n, p_n}$,

$$|f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, Z'_i, \dots, Z_n)| \leq 5\lambda(2p_n)^\alpha.$$

3. $\widehat{R}_{CV} - \widetilde{R}_n$ is closed to zero with high probability

Applying theorem 24, we obtain for all $\varepsilon \geq 0$:

$$\begin{aligned} \Pr(\widehat{R}_{CV} - \widetilde{R}_n \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) &\leq 2(\exp(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^\alpha)^2(1 + \frac{2\varepsilon}{15n(5\lambda(2p_n)^\alpha)})^2}) \\ &\quad + \frac{2n\delta_{n,p_n}'^{1/2}}{5\lambda(2p_n)^\alpha} \exp(\frac{\varepsilon n}{4n(5\lambda(2p_n)^\alpha)^2})) + n\delta_{n,p_n}'^{1/2}) \\ &\leq 2(\exp(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^\alpha)^2(1 + \frac{2\varepsilon}{15n(5\lambda(2p_n)^\alpha)})^2}) \\ &\quad + \frac{2n\delta_{n,p_n}'^{1/2}}{5\lambda(2p_n)^\alpha} \exp(\frac{\varepsilon n}{4n(5\lambda(2p_n)^\alpha)^2})) + n\delta_{n,p_n}'^{1/2}). \end{aligned}$$

By symmetry, we also upper bound $\Pr(\widehat{R}_{CV} - \widetilde{R}_n \leq -(\varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha))$ by the same quantity. \square

Theorem 26 (Weak stability) *Suppose that \mathcal{H} holds. Let Ψ be a machine learning which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$. Then for all $\varepsilon \geq 0$, we have*

$$\Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \lambda(2p_n)^\alpha) \leq 2\exp(-2np_n\varepsilon^2) + \kappa(n)\delta_{n,p_n}$$

where $\kappa(n)$ is the number of elements in the cross-validation.

Furthermore, if the distance is the uniform distance d_∞ , we have

$$\begin{aligned} \Pr(|\widehat{R}_{CV} - \widetilde{R}_n| \geq \varepsilon + \delta_{n,p_n} + \lambda(2p_n)^\alpha) &\leq 4(\exp(-\frac{\varepsilon^2}{10n(5\lambda(2p_n)^\alpha)^2(1 + \frac{2\varepsilon}{15n(5\lambda(2p_n)^\alpha)})^2}) \\ &\quad + \frac{2n\delta_{n,p_n}'^{1/2}}{5\lambda(2p_n)^\alpha} \exp(\frac{\varepsilon n}{4n(5\lambda(2p_n)^\alpha)^2})) + n\delta_{n,p_n}'^{1/2}) \end{aligned}$$

with $\delta_{n,p_n}' = \delta_{n,1/n} + \kappa(n)\delta_{n,p_n}$

Proof.

For the first inequality, it is enough to use remarks 8.

For the second, it is enough to follow the previous proofs and to notice that $\Pr(\cup_{v_n^{tr}} B_{v_n^{tr}}) \leq \kappa(n)\delta_{n,p_n}$. \square

Similar results for hold-out can be derived in the spirit of proposition 22. We can now use the previous probability upper bounds to derive upper bounds for the expectation of $|\widehat{R}_{CV} - \widetilde{R}_n|$.

3.4 Results for the L_1 norm

For the sake of simplicity, we suppose here that $\alpha = 1$. In the general case, we just consider the weakest notion: weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stability.

Theorem 27 (L_1 norm of cross-validation estimate) Suppose that \mathcal{H} holds. Let Ψ be a machine learning which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable. Then, we have

$$\mathbb{E}_{\mathcal{D}_n} |\hat{R}_{CV} - \tilde{R}_n| \leq 2\lambda p_n + \sqrt{\frac{2}{np_n}} + \delta_{n,p_n}.$$

Furthermore, if Ψ is a machine learning which is strong $(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, we have

$$\mathbb{E}_{\mathcal{D}_n} |\hat{R}_{CV} - \tilde{R}_n| \leq \delta_{n,p_n} + 2\lambda p_n + 51\lambda\sqrt{n}p_n + \frac{n}{9\lambda p_n} \delta'_{n,p_n},$$

with $\delta'_{n,p_n} = \delta_{n,p_n} + (n+1)\delta_{n+1,1/n+1}$

Proof.

These inequalities are a consequence of the previous propositions and of the following lemma (for a proof, see e.g. [DGL96]):

Lemma 28 Let X be a nonnegative random variable. Let K, C nonnegative real such that $C \geq 1$. Suppose that for all $\varepsilon > 0$, $\mathbb{P}(X \geq \varepsilon) \leq C \exp(-K\varepsilon^2)$. Then:

$$\mathbb{E}X \leq \sqrt{\frac{\ln(C) + 2}{K}}.$$

For the second one, it is enough to follow the previous proofs and to notice that $\Pr(\cup_{V_n^{tr}} B_{V_n^{tr}}) \leq \kappa(n)\delta_{n,p_n}$

□

We deduce that

Corollary 29 Suppose that \mathcal{H} holds. If Ψ be a machine learning which is weak $(\lambda, (\delta_{n,p_n})_{n,p_n}, d)$ stable, we define the splitting rule $p_n^* = (1/\sqrt{24}\lambda)^{2/3}(1/n)^{1/3}$. Then, we have

$$\mathbb{E}_{\mathcal{D}_n} |\hat{R}_{CV} - \tilde{R}_n| \leq 4(\lambda/n)^{1/3}.$$

Furthermore, if Ψ is a machine learning which is strong $(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, we use leave-one-out cross-validation for n large enough. And we have

$$\mathbb{E}_{\mathcal{D}_n} |\hat{R}_{CV} - \tilde{R}_n| = O_n(\lambda/\sqrt{n}).$$

Proof.

Recall that for a large class of learning algorithm, we have in mind that $\delta_{n,p_n} = O_n(p_n \exp(-n(1-p_n)))$. Thus $2\lambda p_n + \sqrt{\frac{2}{np_n}} + \delta_{n,p_n} \leq 4\lambda p_n + \sqrt{\frac{2}{np_n}}$. We can differentiate this last bound seen as a function of p_n . We obtain $p_n^* = (1/\sqrt{24}\lambda)^{2/3}(1/n)^{1/3}$. Thus, we deduce that $\mathbb{E}_{\mathcal{D}_n} |\hat{R}_{CV} - \tilde{R}_n| \leq 4(\lambda/n)^{1/3}$. If Ψ is a machine learning which is strong $(\lambda, (\delta_n)_n, d_\infty, \mathbb{Q})$ stable, we obtain $\delta_{n,p_n} + 2\lambda p_n + 51\lambda\sqrt{n}p_n + \frac{n}{9\lambda p_n} \delta'_{n,p_n} \leq 4\lambda p_n + 51\lambda\sqrt{n}p_n$ for n large enough since $\frac{n}{9\lambda p_n} \delta'_{n,p_n} = O_n(n^3 \exp(-n/2))$ if $p_n \leq 1/2$. Thus, $p_n^* = 1/n$ for n large enough and $\mathbb{E}_{\mathcal{D}_n} |\hat{R}_{CV} - \tilde{R}_n| = O_n(\lambda/\sqrt{n})$.

□

We have obtained the following conclusions:

- Cross-validation is consistent as an estimator of the generalization error of stable algorithms.
- There is a tradeoff interpretation in the choice of the proportion of elements p_n of the test set: the smaller p_n is, the greater the term $B(n, p_n, \varepsilon)$ is controlled but the less the term $V(n, p_n, \varepsilon)$ is upper bounded.

- In the general setting, our bounds require that the sizes of the training set and the test set grow to infinity.
- In the particular case of the stability with respect to the most stable kind of stability (namely the uniform stability), we can have a stronger result: the number of elements in the test set does need to grow to infinity for the consistency of symmetric cross-validation procedures. But we lose this property with the hold-out cross-validation.
- Symmetric cross-validation out performs hold-out cross-validation for large sets.
- At last, as far as the expectation $\mathbb{E}_{\mathcal{D}_n} |\hat{R}_{CV} - \tilde{R}_n|$ is concerned, we can define a splitting rule in the general setting.

References

- [AL68] D. M. Allen, The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 1968, 16, 125-127.
- [ATE92] M. Atteia. Hilbertian kernels and spline functions. North-Holland, 1992.
- [ARL07] S. Arlot, Model selection by resampling penalization. *submitted to COLT* 2007.
- [BEN04] Y. Bengio and Y. Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research* 5, 1089-1105, 2004.
- [BIS05] M. Markatou, H. Tian, S. Biswas, G. Hripcsak. Analysis of Variance of Cross-Validation Estimators of the Generalization Error. *Journal of Machine Learning Research* 1127-1168, 2005.
- [BREI84] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. Classification and regression trees. *The Wadsworth statistics probability series*. Wadsworth International Group, 1984.
- [BREI92] L. Breiman, and Spector, P. (1992), Submodel selection and evaluation in regression: The X-random case *International Statistical Review*, 60, 291-319.
- [BREI96] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140.
- [BKL99] A. Blum, A., Kalai, A., and Langford, J. (1999). Beating the hold-out: Bounds for k-fold and progressive cross-validation. *Proceedings of the International Conference on Computational Learning Theory*.
- [BE01] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance *In Advances in Neural Information Processing Systems* 13: Proc. NIPS’2000, 2001.
- [BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2002.
- [BUR89] P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76:503– 514, 1989.
- [BTW07] F. Bunea, A.B. Tsybakov and M.H. Wegkamp, M. H. Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, 1 169?194, 2007.
- [COR09] M.Cornec. Concentration inequalities of the cross-validation estimator for Empirical Risk Minimiser. Technical Report. 2009.
- [DGL96] L. Devroye, L. Györfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Number 31 in *Applications of Mathematics*. Springer, 1996.
- [DW79] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory*, 25(5):601 604, 1979. 41
- [DEWA79] L. P. Devroye and T. J. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, IT?25(2):202?207, 1979
- [DUD03] S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in model selection and performance assessment. *Technical Report* 126, Division of Biostatistics, University of California, Berkeley, 2003.
- [DUD04] S. Dudoit, M. J. van der Laan, S. Keles, A. M. Molinaro, S. E. Sinisi, and S. L. Teng. Loss-based estimation with cross-validation: Applications to microarray data analysis. *SIGKDD Explorations, Microarray Data Mining Special Issue*, 2004.

- [DUD04BIS] M.J. van der Laan, S. Dudoit, A. van der Vaart (2004), The cross-validated adaptive epsilon-net estimator, submitted for publication in *Statistics and Decisions*.
- [FRE95] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. of the Second European Conference on Computational Learning Theory*. LNCS, March 1995.
- [GEI75] S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975.
- [GYO02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002a.
- [HTF01] T. Hastie, R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2001.
- [HOEF63] W. Hoeffding, (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58, 13?30.
- [HOL96] S. B. Holden. Cross-validation and the PAC learning model. *Research Note* RN/96/64, Dept. of CS, Univ. College, London, 1996.
- [HOL96bis] S. B. Holden. PAC-like upper bounds for the sample complexity of leave-one-out cross validation. In *Proceedings of the Ninth Annual ACM Workshop on Computational Learning Theory*, pages 41–50, 1996.
- [KR99] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11:1427–1453, 1999.
- [KEA95] M. Kearns, (1995). A bound on the error of cross validation, with consequences for the training-test split. In *Advances in Neural Information Processing Systems 8*. The MIT Press.
- [KMNR95] M. J. Kearns, Y. Mansour, A. Ng., and D. Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth Annual ACM Workshop on Computational Learning Theory*, pages 21–30, 1995. To Appear in Machine Learning, COLT95 Special Issue.
- [KUT02] S. Kutin. Extensions to McDiarmid’s inequality when differences are bounded with high probability. *Technical report*, Department of Computer Science, The University of Chicago, 2002. In preparation.
- [KUNIY02] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error, 2002. *Technical report* TR-2002-03, University of Chicago.
- [KUNIY01] S. Kutin and P. Niyogi. The interaction of stability and weakness in AdaBoost. *Technical Report* TR-2001-30, Department of Computer Science, The University of Chicago, 2001.
- [LM68] P. A. Lachenbruch,; M. Mickey, Estimation of error rates in discriminant analysis. *Technometrics* 1968, 10, 1-11.
- [Li87] K-C Li. Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15:958–975, 1987.
- [Lug03] G Lugosi. Concentration-of-measure inequalities presented at *the Machine Learning Summer School 2003*, Australian National University, Canberra,
- [McC76] P. J. McCarthy. The use of balanced half-sample replication in crossvalidation studies. *Journal of the American Statistical Association*, 71: 596–604, 1976.

- [McD89] C. McDiarmid. On the method of bounded differences. *In Surveys in combinatorics*, 1989 (Norwich, 1989), pages 148–188. Cambridge Univ. Press, Cambridge, 1989.
- [McD98] C. McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, Berlin, 1998.
- [PIC84] R. R. Picard and R. D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79:575–583, 1984.
- [RIP96] B. D. Ripley. Pattern recognition and neural networks. *Cambridge University Press*, Cambridge, New York, 1996.
- [ROG63] C. Rogers. Covering a sphere with spheres. *Mathematika*, vol. 10, pp. 157–164, 1963.
- [SHAO93] J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993. !
- [STO74] M. Stone, (1974). Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36, 111–147.
- [STO77] M. Stone, (1977). Asymptotics for and against cross-validation. *Biometrika*, 64, 29–35.
- [VAL84] L.G. Valiant (1984). A theory of learnable. *Proc. of the 1984, STOC*, pages 436–445.
- [Vaart96] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.
- [VA71] V. Vapnik, and A. Chervonenkis, (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16, 264–280.
- [VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.
- [VA82] V. Vapnik, (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag.
- [Vap95] V. Vapnik. *The nature of statistical learning theory*. Springer, 1995.
- [Vap98] V. Vapnik. *Statistical learning theory*. John Wiley and Sons Inc., New York, 1998. A Wiley-Interscience Publication.
- [YAN07] Y. Yang, Consistency of Cross Validation for Comparing Regression Procedures. *Accepted by Annals of Statistics*.
- [ZHA93] P. Zhang. Model selection via multifold cross-validation. *Annals of Statistics*, 21:299–313, 1993.
- [ZHA00] T. Zhang. A leave-one-out cross validation bound for kernel methods with applications in learning. *14th Annual Conference on Computational Learning Theory*, 2001 - Springer.

4 Appendices

4.1 Inequalities

We recall three very useful results. The first one, due to [HOEF63], bounds the difference between the empirical mean and the expected value. The second one, due to [VC71], bounds the supremum over the class of predictors of the difference between the training error and the generalization error. The last one is called the bounded differences inequality [McD89].

Theorem 30 (Hoeffding's inequality) *Let X_1, \dots, X_n independent random variables in $[a_i, b_i]$. Then for all $\varepsilon > 0$, we get*

$$\mathbb{P}(\sum X_i - \mathbb{E}(\sum X_i) \geq n\varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_i (b_i - a_i)^2}}.$$

Theorem 31 (McDiarmid, [McD89]) *Let X_1, \dots, X_n be independent random variables taking values in a set A , and assume that $f : A^n \rightarrow \mathcal{R}$ satisfies*

$$\forall i, \sup_{\substack{x_1, \dots, x_i, \dots, x_n \\ x_i'}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i'}, \dots, x_n)| \leq c_i.$$

Then for all $\varepsilon > 0$, we have

$$\mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n) \geq \varepsilon) \leq e^{-\frac{2\varepsilon^2}{\sum_i c_i^2}}.$$